

RESEARCH ARTICLE

A Novel Decision Level Class-Wise Ensemble Method in Deep Learning for Automatic Multi-Class Classification of HER2 Breast Cancer Hematoxylin-Eosin Images

PATEEL G. P. , KEDARNATH SENAPATI , AND ABHISHEK KUMAR PANDEY 

Department of Mathematical and Computational Sciences, National Institute of Technology Karnataka, Surathkal, Mangaluru 575025, India

Corresponding author: Kedarnath Senapati (kedar@nitk.edu.in)


ABSTRACT The Human Epidermal Growth Factor Receptor 2 (HER2) is one of the aggressive subtypes of breast cancer. The HER2 status decides the requirement of breast cancer patients to receive HER2-targeted therapy. The HER2 testing involves combining Immunohistochemistry (IHC) screening, followed by fluorescence in situ hybridization (FISH) for cases where IHC results are equivocal. These tests may involve multiple trials, are time intensive, and tend to be more expensive for certain classes of people. Hematoxylin and Eosin (HE) staining is employed for visualizing general tissue morphology and is a routine, cost-effective method. In this study, we introduce a novel automated class-wise weighted average ensemble deep learning algorithm at the decision level. The proposed algorithm fuses three pre-trained deep-learning models at the decision level by assigning a weight to each class based on their performance of the model to classify the HE-stained breast histopathology images into multi-class HER2 statuses as HER2-0+, HER2-1+, HER2-2+, and HER2-3+. The class-wise weight allocation to the base classifiers is one of the key features of the proposed algorithm. The presented framework surpasses all the existing methods currently employed on the Breast Cancer Immunohistochemistry (BCI) dataset, establishing itself as a dependable approach for detecting HER2 status from HE-stained images. This study highlights the robustness of the proposed algorithm as well as the sufficient information encapsulated within HE-stained images for the effective detection of the HER2 protein present in breast cancer. Therefore, the proposed method possesses the potential to sideline the need for IHC laboratory tests, which hoard time and money.

INDEX TERMS HER2, ensemble learning, histopathology, breast cancer, multi-class classification.

I. INTRODUCTION

Breast cancer stands as the most common cancer among women worldwide. It is a neoplasm characterized by significant heterogeneity, encompassing distinct subtypes. The worldwide occurrence of breast cancer has been increasing, and there is a projected 46% rise in cases by 2040 [1], [2]. The improvement in survival rates can be attributed to significant advancements in screening methods, early diagnosis, and breakthroughs in treatment options [3]. HER2-positive breast cancer is characterized by a high

degree of malignancy. It is a distinct subtype known for its aggressive behavior, early recurrence, metastatic potential, and poor prognosis. HER2 expression is present in around 15–25% of breast cancers, and its status plays a crucial role in determining the most suitable treatment required [4], [5], [6]. HER2 over-expression is among the initial events in breast carcinogenesis. The presence of HER2 protein increases the rate of metastatic and recurrent breast cancers by 50%, and in some cases, even up to 80% [5]. The most recent guidelines emphasize the importance of undergoing regular HER2 testing for patients diagnosed with invasive breast cancer, recurrent, and metastatic tumors [7].

The associate editor coordinating the review of this manuscript and approving it for publication was Szidonia Lefkovits .

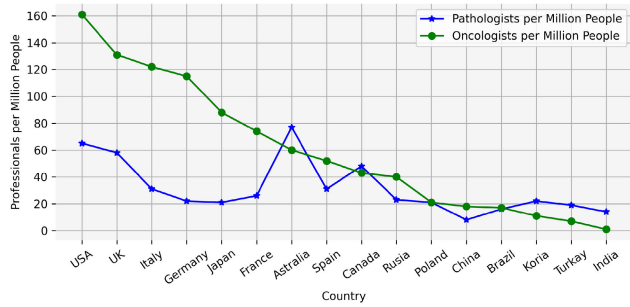


FIGURE 1. Pathologists and oncologists availability across selected countries (density per million population).

TABLE 1. IHC scores indicating HER2 outcome by american cancer society.

IHC Score	HER2 Status	Treatment Plan
0	HER2-Negative	No response for HER2 targeting treatment
1	HER2-Negative	Advanced research may help in certain instances
2	Equivocal	Further test is necessary with FISH to clarify
3	HER2-Positive	These cancers are usually treated with drugs that target HER2

Detecting breast malignancy early improves survival rates significantly. Pathologists typically use conventional methods like HE staining to examine crucial morphological features such as shapes, patterns, and structures of cells and tissue essential for cancer diagnosis. IHC is another staining method employed to validate the existence of various breast cancer subtypes. This technique utilizes antibodies to identify different antigens, including HER2, Progesterone Receptor (PR), and Estrogen Receptor (ER) [8]. Outcomes of IHC staining are classified into various HER2 status scores ranging from 0 to 3+. A score of 0 or 1+ signifies HER2-negative, and 3+ is HER2-positive. However, a score of 2+ requires further testing through FISH to determine the HER2 gene status [9]. The precise evaluation of the HER2 receptor is crucial in identifying the types of breast cancer patients requiring therapy targeting the HER2 antigen [10]. However, the shortage of pathologists, oncologists as well as diagnostic facilities worldwide deprives many needy people of the opportunity to consult them in time [11], [12]. The Fig. 1 illustrates the distribution of pathologists and oncologists density (per million population) across various countries.

In recent times, deep convolution neural networks (CNN) have gained widespread recognition as highly potent tools for image classification. They possess distinct advantages over conventional machine learning methods by offering an end-to-end framework for feature extraction and classification. This framework eliminates the need for users to engage in laborious handcrafted feature extraction, enhancing

efficiency and effectiveness [13], [14]. The accomplishments of deep CNNs have led numerous researchers to adopt these methods for classifying histopathological microscopic images. Despite its widespread use, the single deep CNN model has constraints in extracting discriminate features, potentially leading to sub-optimal classification performance. As a solution, an ensemble of deep CNN architectures has been developed to capture the representation of histopathological microscopic images from diverse perspectives, aiming to achieve more accurate classification results [15], [16].

Cutting-edge and enhanced deep learning methodologies, currently under development, to identify early-stage breast malignancies with HER2 status are precisely using histopathological HE-stained images [17], [18]. These methodologies are designed to support pathologist’s decision-making process. A substantial dataset is essential for training a deep learning model, however, adequate number of medical images in the public domain are not easily available. Therefore, working with available smaller dataset for research purposes, transfer learning techniques can reduce the training time and enhance model performance. Any of the following three approaches can be used for transfer learning. In the first approach, a pre-trained CNN model can serve as a feature extractor in the architecture. The second approach involves the fine-tuning of hyperparameters as well as, the weights of the final layer only, in a pre-trained CNN are modified. The third technique makes similar modifications to the entire architecture [8].

Most of the earlier investigations have revolved around the binary classification of the HER2 subtype. In this case, the straightforward binary classification task for HER2 breast histopathology images can be misleading for cancer professionals and might not provide adequate information for determining the treatment plan. Also, limited attention has been given to predicting HER2 breast carcinoma from HE-stained images, particularly in multi-stage classification. Addressing this challenge can bypass the IHC staining laboratory test and also reduce the diagnostic time as well as the cost involved in the test. The American Cancer Society recommends HER2 testing, either through biopsy specimens or the surgical removal of the tumor for all invasive breast cancers. The outcome of the HER2 test is shown in Table 1, which provides extra information to professionals for treatment.

This motivation prompts us to focus on and tackle the challenges associated with the multi-class classification of HER2 breast cancer using HE-stained images. In this paper, we have proposed a decision-level class-wise weighted average ensemble learning algorithm by fusing the three chosen base classifiers GoogLeNet [19], WideResNet-50 [20], and DenseNet-201 [21]. The framework of the proposed work is as follows. First, we extract multiple-scale patches of size 256×256 , 512×512 , and 1024×1024 from the HE-stained images. Then, the extracted patches of different scales are fed into the base classifiers separately, out of which the best-performing scale is selected for further study.

The selected patches are used to fine-tune the chosen base classifiers. Finally, the fine-tuned three base classifiers are fused through the proposed algorithm. We emphasize that the fine-tuning of multiple base classifiers has the potential to extract diverse and complementary visual features, enabling a more comprehensive representation of images from different perspectives.

The rest of the paper is organized as follows. In Section II, we discuss about the works related to our context. The acquisition of the dataset and the proposed method are presented in Section III. The performance of the proposed model is evaluated, analyzed, and listed in Section IV. Section V discusses some future scopes of the present paper along with a conclusion.

II. RELATED WORKS

The rise and evolution of deep learning have been assisting in numerous breakthroughs across various computer vision applications, encompassing the classification of natural and medical images [22]. Several highly effective CNN architectures, including AlexNet [23], VGG [24], GoogLeNet [19], WideResNet-50 [20], and DenseNet-201 [21], were developed specifically for the ILSVRC ImageNet classification task [25]. Notably, ResNet, DenseNet and GoogLeNet stand out as the most widely adopted network backbone structures, exhibiting superior performance compared to alternative machine learning approaches.

In [26], Oliveira et al. introduced a CNN model that integrated multiple instance learning techniques to ascertain HER2 status from HE images. The CNN model was initially pre-trained using IHC images from the HER2SC dataset. Subsequently, fine-tuning was performed using HE images from the same dataset. The model was then tested on HE-stained slides from the CIA-TCGA-BRCA (BRCA) dataset, achieving test accuracies of 83.3% and 53.8% for the respective datasets.

In [27], Maleki et al. introduced a method to improve the speed and precision of histopathology image classification as Benign and Malignant. The process involved feature extraction utilizing transfer learning models, with subsequent classification performed using Extreme Gradient Boosting (XGBoost). The obtained accuracy rates were 93.6% at 40X magnification, 91.3% at 100X magnification, 93.8% at 200X magnification, and 89.1% at 400X magnification.

In another study, Shamai et al. [28] conducted the prediction of three biomarkers, namely, ER, PR, and HER2 utilizing the ResNet architecture. The outcomes of this investigation revealed AUC values of 80%, 75%, and 74% for the respective three biomarkers.

Shovon et al. [29] introduced a deep transfer learning model based on a modified Xception network for the multistage classification of HER2 from HE images, utilizing the BCI dataset. The model achieved an accuracy of 87.01%. In a different approach, DenseNet-201 and Xception were combined as a single classifier in an ensemble, utilizing feature confidence scores to establish the decision boundary

and achieving an impressive accuracy of 97.12% [30]. Mridha et al. presented a unique CNN model in [31] to classify multistage HER2 from HE images, achieving an accuracy of 85.10%.

A comprehensive ensemble model was proposed in [32], leveraging image-level annotations for binary classification of breast histopathology images, distinguishing benign and malignant lesions. The ensemble network, which included VGG-16, Xception, ResNet-50, and DenseNet-201, used accuracy only, as the weighting factor, resulting in a binary image-level classification accuracy of 98.90%.

III. MATERIALS AND METHODS

In this section, we discuss the dataset used for this work as well as the proposed method, which consists of three steps: (1) Patch extraction and scale selection, (2) Training deep CNN models, and (3) Class-wise ensemble method. The diagram that describes the proposed workflow is shown in Fig. 5.

A. DATASET

This study utilizes a recently released benchmark dataset for breast cancer (BC) immunohistochemical analysis, known as the BCI dataset [33]. The dataset was created using the Hamamatsu NanoZommer S60 scanner with a resolution of $0.46 \mu\text{m}$ per pixel. 600 whole-slide imaging (WSI) slides were scanned, each consisting of 20,000 pixels. These slides were split into 16 blocks, each size 1024×1024 pixels per patch. The resulting dataset comprises 4870 pairs of HE and IHC images, each with dimension 1024×1024 . This dataset is labeled by medical experts into four distinct classes denoted as 0+, 1+, 2+, and 3+, as illustrated in Fig. 2. The multistage HER2 status labeling of HE images was performed using the corresponding IHC image scores. To the best of our knowledge, the BCI dataset is the only publicly available dataset with four-class HER2 status as per the CAP/ASCO guidelines [9].

B. PATCH EXTRACTION AND SCALE SELECTION

BCI HE images of size 1024×1024 are considered as scale-1. The non-overlapping patches of sizes 512×512 and 256×256 extracted separately from the scale-1 are considered scale-2 and scale-3 respectively. The extracted image patches with a grey limit of more than 0.8 are discarded as they do not contain any tissue information. Each scale images are separately fed to the chosen three base classifiers as shown in Fig. 3. Among the three different scale images, scale-2 gives better accuracy and is selected for further study, which we term as the 'scale selection procedure'. Then, the selected scale-2 images are split into training, validation, and testing sets as shown in Table 2.

C. TRAINING DEEP CNN MODELS

Considering their exceptional performance in image classification, GoogLeNet [19], WideResNet-50 [20], and

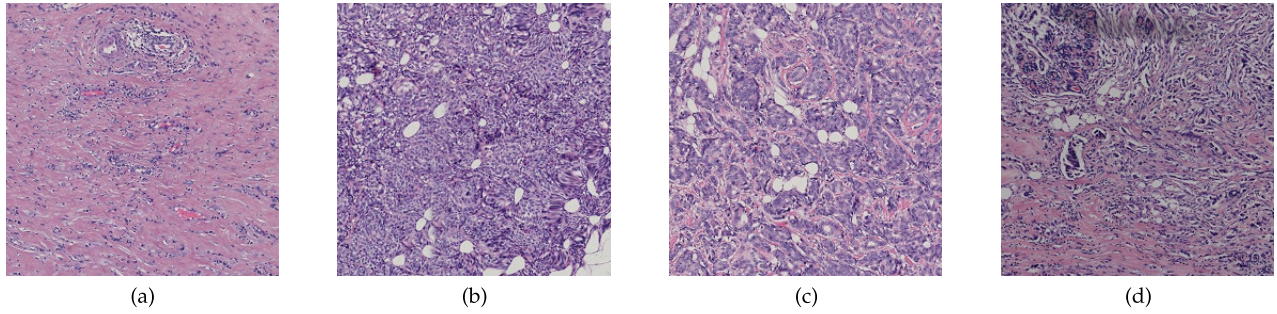


FIGURE 2. HE image samples of BCI dataset (a) HER2-0+; (b) HER2-1+; (c) HER2-2+; (d) HER2-3+.

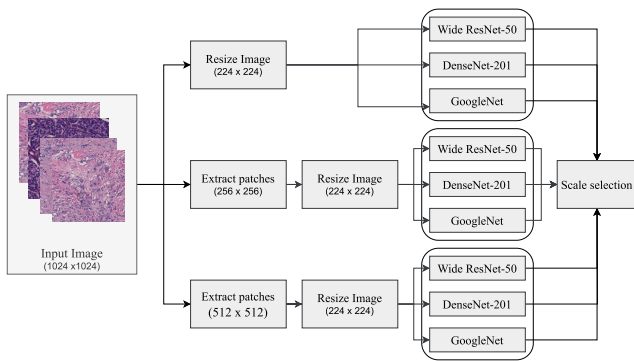


FIGURE 3. Scale selection process.

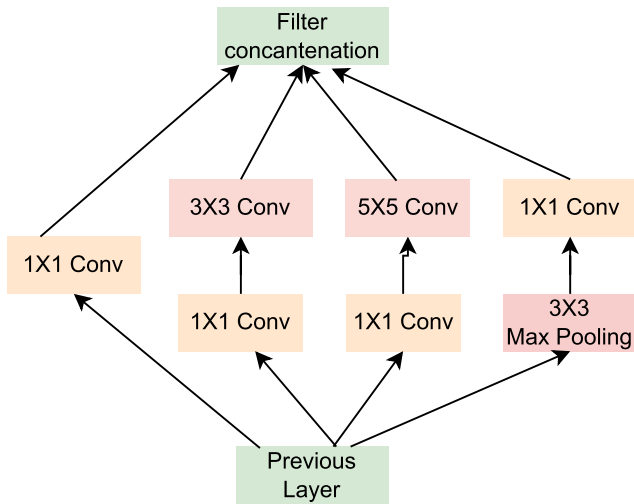


FIGURE 4. GoogleNet inception block.

DenseNet-201 [21] have been chosen as the backbone networks for this study.

GoogleNet, also known as Inception-v1, was developed to address the challenges associated with training extremely deep neural networks. One of the notable features of GoogleNet is the integration of “Inception” modules, deploying multiple parallel convolution filters of varying sizes within a layer, as illustrated in Fig. 4. This innovative approach allows the network to capture features across diverse scales and resolutions simultaneously. The strategic design of GoogleNet empowers it to achieve

TABLE 2. BCI HE image distributions after preprocessing.

	HER2-0+	HER2-1+	HER2-2+	HER2-3+	Total
Train	609	3011	5385	3333	12338
Validate	624	405	777	472	2278
Test	121	804	1562	899	3386
					18002

remarkable accuracy while maintaining computational efficiency.

The DenseNet architecture was developed to address specific limitations inherent in traditional deep neural networks, including issues like vanishing gradients and constraints on efficiently reusing features. In the design of DenseNet, each layer not only depends on the input from the previous layer but also establishes direct connections with all other preceding layers. This interconnected structure within the dense block facilitates the network to enhance information dissemination and proactively encourages the re-utilization of features.

The Wide-ResNet-50 architecture was developed to address the issue of training deep residual networks, which posed a challenge of diminishing feature reuse, leading to a significant slowdown in the training process. The decrease in depth and increase in the width of residual networks demonstrate superior performance compared to their commonly used very deep and thin counterparts.

In the process of customizing each pre-trained deep CNN model for a four-category HER2 image classification task, we retain four neurons in the last fully connected layer, excluding the remaining output neurons along with their associated weights. Consequently, during the fine-tuning of pre-trained deep CNN models, the scale-2 images that are selected in the scale selection process discussed in Subsection III-B are given as input. The outputs generated are four-dimensional vectors, indicating the probabilities of input images belonging to each of the HER2-0+, HER2-1+, HER2-2+, and HER2-3+ status.

Due to the limited number of training histopathology HE images, each deep CNN model underwent pre-training and convergence on the ImageNet challenge dataset. We utilized

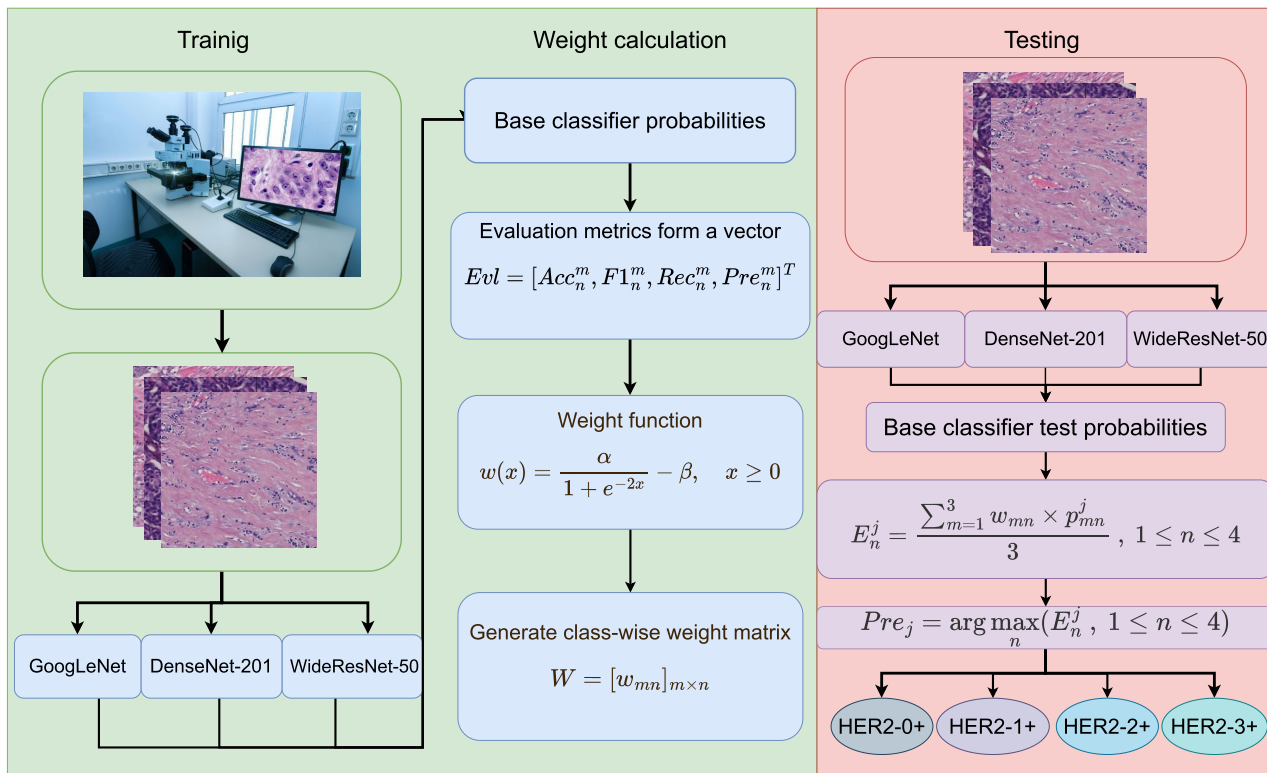


FIGURE 5. Architecture of the proposed method; x represents evaluation metrics score; p_{mn}^j and w_{mn} respectively represent the probability score of j^{th} test image and weight assigned to the n^{th} class of m^{th} base classifier.

pre-trained models available from PyTorch. For the fine-tuning of each deep CNN model, we employed the SGD optimizer to minimize the cross-entropy loss, setting the number of epochs and batch size to 100 and 20 respectively, and the learning rate to 0.001 with a decay of 10% in every 20 training epochs.

D. PROPOSED CLASS-WISE ENSEMBLE METHOD

Ensemble learning entails the integration of multiple learning algorithms, resulting in a robust and dependable model with improved generalisability. These methods use the strength of multiple models, effectively address the limitations of individual models, and provide predictions with higher accuracy. Different ensemble techniques used by numerous authors have proven their efficacy for various deep learning tasks. Out of these, the weighted average ensemble method turns out to be a powerful technique for combining different classifiers. The effectiveness of this method relies on the careful assignment of weights to the individual base classifiers.

In the conventional weighted average ensemble method for multi-class classification tasks, the fusion of multiple base classifiers involves multiplying all the class probabilities with the calculated weight assigned to each individual base classifier. However, this approach assumes equal performance of the base classifier across all classes. In reality, the base classifier may not perform equally on all classes, leading to misclassification. To address this issue, we propose a

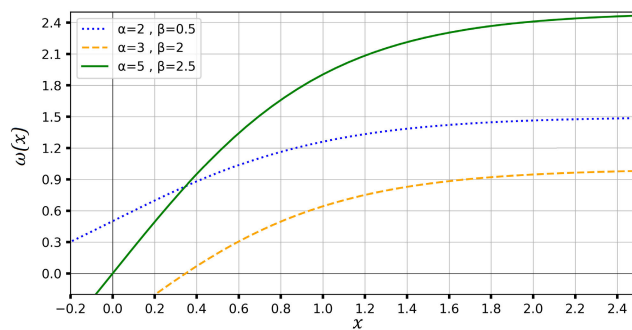


FIGURE 6. Graph of the weight allocation function $w(x)$ used in the proposed work.

decision-level, class-wise, weighted average ensemble deep learning algorithm. Here, the contribution of each class of all the base classifiers to the final classification is weighted based on the base classifier’s performance in each class. This weight allocation procedure is discussed below and is outlined in Algorithm 1.

1) TRAINING AND WEIGHT CALCULATION

First, the probability scores of both the training and validation set obtained during the training phase are utilized to calculate the evaluation metrics, which form a vector as follows

$$Evl = [Acc_n^m, F1_n^m, Rec_n^m, Pre_n^m]^T \tag{1}$$

Algorithm 1 Class-wise ensemble algorithm

Input:
 $D \leftarrow$ Dataset
 $\alpha \leftarrow$ learning rate
 $optimizer \leftarrow SGD$
 $e \leftarrow$ Number of epochs
 $Models \leftarrow$
 [DenseNet-201, GoogLeNet, WideResNet-50]

Output:
 $w_{mn} \leftarrow m^{th}$ base learner n^{th} class weight
 $p_{mn}^j \leftarrow$
 j^{th} test image probability of m^{th} base learner n^{th} class
 Predicted Class of test image

Training Phase:
for $model \leftarrow Models$ **do**
 Initialize the training and validation set for model
 Initialize the hyper-parameters of the model **while**
 $epochs \leq e$ **do**
 | fine-tune the pre-trained model
end
 store the probability scores
 of both training and validation set
end
 Compute the class-wise accuracy, Precision,
 Recall, F1-score, and form a vector
 Evl

Function ClaculateWeights (Evl):
 $w(x) = 0, W = []$
for $m \leftarrow Models$ **do**
for $n \leftarrow Evl$ **do**
for $x \leftarrow$ parameters of Evl vector **do**
 | $w(x) = w(x) + \frac{\alpha}{1+e^{-2x}} - \beta$
end
 $w_{mn} = w(x)/4$
 $W = [W, w_{mn}]$
end
end
return W

Test Phase:
for $images \leftarrow testset$ **do**
 compute and store the
 test probability scores
end
 p_{mn}^j

Function EnsembleProb (W, p_{mn}^j):
 $E = 0, E_n^j = []$
for $n \leftarrow W$ **do**
for $m \leftarrow model$ **do**
 | $E = E + (w_{mn} \times p_{mn}^j)$
end
 $E = E/3$
 $E_n^j = [E_n^j, E]$
end
return ($E_n^j, 1 \leq n \leq 4$)

Prediction:
 $Pre_j = \arg \max(E_n^j, 1 \leq n \leq 4)$

represents the number of classes of base classifiers. And, $Acc_n^m, F1_n^m, Rec_n^m,$ and Pre_n^m are respectively the accuracy, F1-score, recall, and precision of n^{th} class for the m^{th} base classifiers.

As the range of evaluation metrics parameters is $[0, 1]$, we choose a suitable weight function as follows

$$w(x) = \frac{\alpha}{1 + e^{-2x}} - \beta, \quad x \geq 0 \quad (2)$$

where the shifting and the scaling parameters $\alpha = 5$ and $\beta = 2.5$, respectively, are selected experimentally for effective allocation of weight to the classes of the base classifier.

The calculated evaluation metrics are utilized to assign the weight for each class of the base classifiers using the weight function $w(x)$, which is plotted in Fig. 6.

One can observe from Fig. 6 that, $w(x) \in [0, 1.9]$ for $x \in [0, 1]$, where x is one of the components of the vector Evl , based on which the weights are assigned to the classes of base classifiers. A smaller value of x results in the weight function generating a smaller weight, thereby giving low priority to the corresponding class of the base classifiers. Conversely, when the value of x is high, the weight function generates a more significant weight, thereby giving high priority to the corresponding class of the base classifiers.

Therefore, the weight w_{mn} of n^{th} class and m^{th} base classifiers is formulated as,

$$w_{mn} = \frac{w(Acc_n^m) + w(F1_n^m) + w(Rec_n^m) + w(Pre_n^m)}{4} \quad (3)$$

and all such weights form a matrix $W = [w_{mn}]_{m \times n}$. Here, $w(Acc_n^m), w(F1_n^m), w(Rec_n^m),$ and $w(Pre_n^m)$ are the weights computed by passing the values of accuracy, F1-score, recall, and precision of n^{th} class of m^{th} base classifier through the weight function $w(x)$ in Eq. 2, for $1 \leq m \leq 3$ $1 \leq n \leq 4$.

2) WEIGHT ALLOCATION AND TESTING

In the subsequent test phase, the test images are provided as input to the trained base classifiers to acquire the test probability scores. The calculated class-wise weights are then multiplied by the test probability scores to yield the weighted average ensemble probability scores, expressed as follows

$$E_n^j = \frac{\sum_{m=1}^3 w_{mn} \times p_{mn}^j}{3} \quad 1 \leq n \leq 4 \quad (4)$$

where, p_{mn}^j represents the probability score of the j^{th} test image for the and n^{th} class and m^{th} base classifiers, E_n^j represents weighted average ensemble probability score for n^{th} class of j^{th} test image.

Finally, the prediction of the class is computed from the weighted average ensemble probability score obtained above using the following expression

$$Pre_j = \arg \max(E_n^j, 1 \leq n \leq 4) \quad (5)$$

where, Pre_j represents predicted class of the j^{th} test image. In this way, the class with the highest probability is computed

where, T denotes transpose, $1 \leq m \leq 3$ represents the number of base classifiers and $1 \leq n \leq 4$

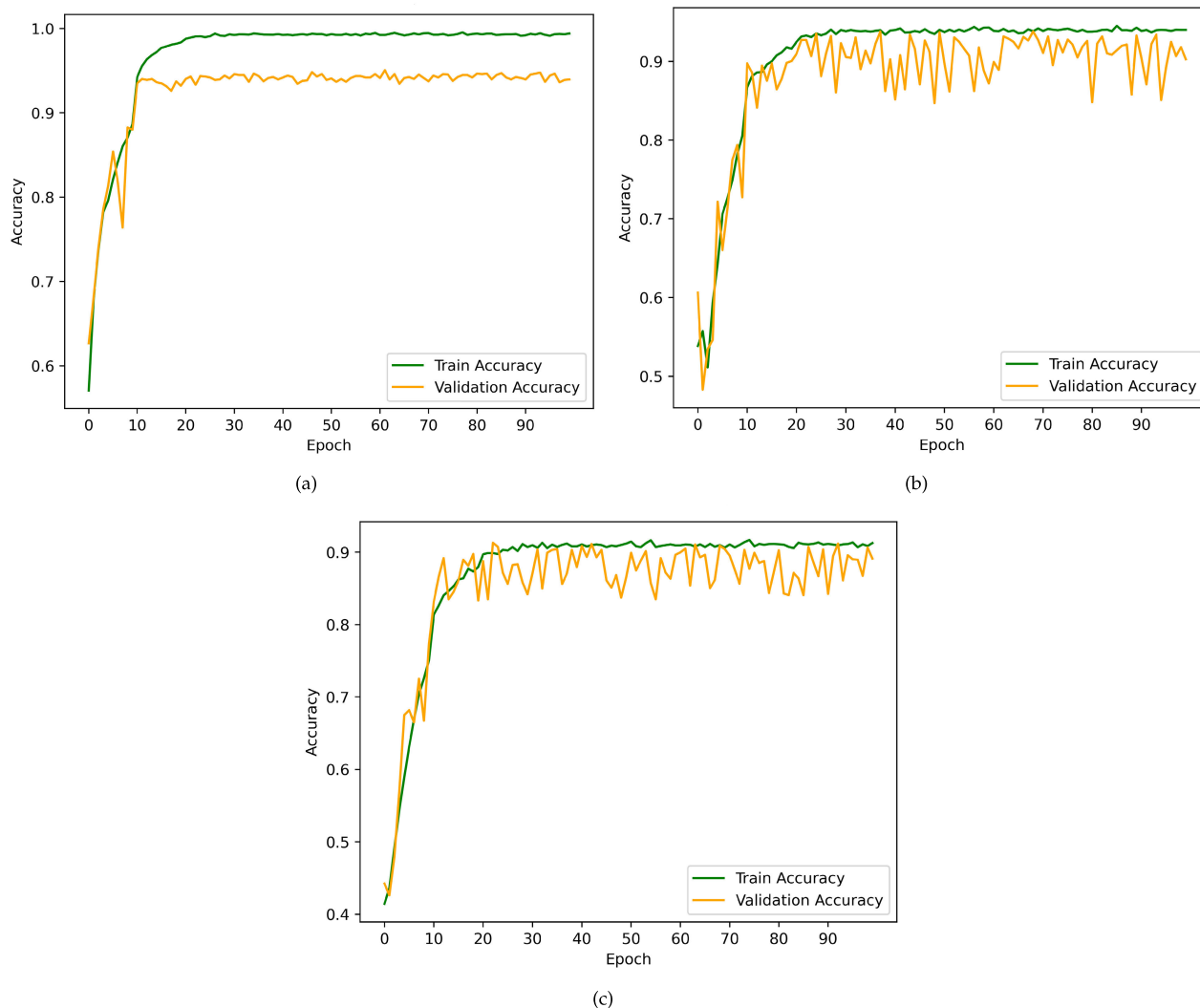


FIGURE 7. Accuracy curve of base classifiers: (a) GoogleNet (b) DensNet-201 (c) WideResNet-50.

TABLE 3. Classification results generated by base classifiers and proposed class-wise ensemble method.

Model	Acc	Pre	Rec	F1	AUC
GoogleNet	95.32	94.12	94.46	94.06	99.00
DenseNet201	94.04	91.53	92.63	92.07	98.00
WideResNet50	92.77	90.89	93.35	91.97	98.00
Proposed Method	97.84	96.62	97.87	97.22	100

as the predicted class of the test image. This predicted class is then compared to the actual label to generate the confusion matrix. Subsequently, accuracy, F1-score, precision, and recall values are calculated based on the confusion matrix.

IV. RESULTS AND DISCUSSION

This section presents the discussion and analysis of the proposed method. Additionally, we compare the presented

TABLE 4. Comparison of proposed class-wise ensemble method with the other traditionally employed ensemble methods in the literature.

Ensemble Method	Acc	Pre	Rec	F1
Average probability	94.98	93.09	95.32	94.11
Weighted Avg Probability	96.40	93.62	96.39	95.27
Majority Voting	94.45	90.39	94.73	92.20
Proposed Method	97.84	96.62	97.87	97.22

model with existing models in the literature to validate the effectiveness of the proposed model.

A. EVALUATION MEASURES

In order to evaluate the performance of the proposed method some measures need to be devised. In this study, we have utilized the confusion matrix to investigate a range of evaluation metrics. The accuracy, precision, recall, and F1-score are the popularly used performance measures,

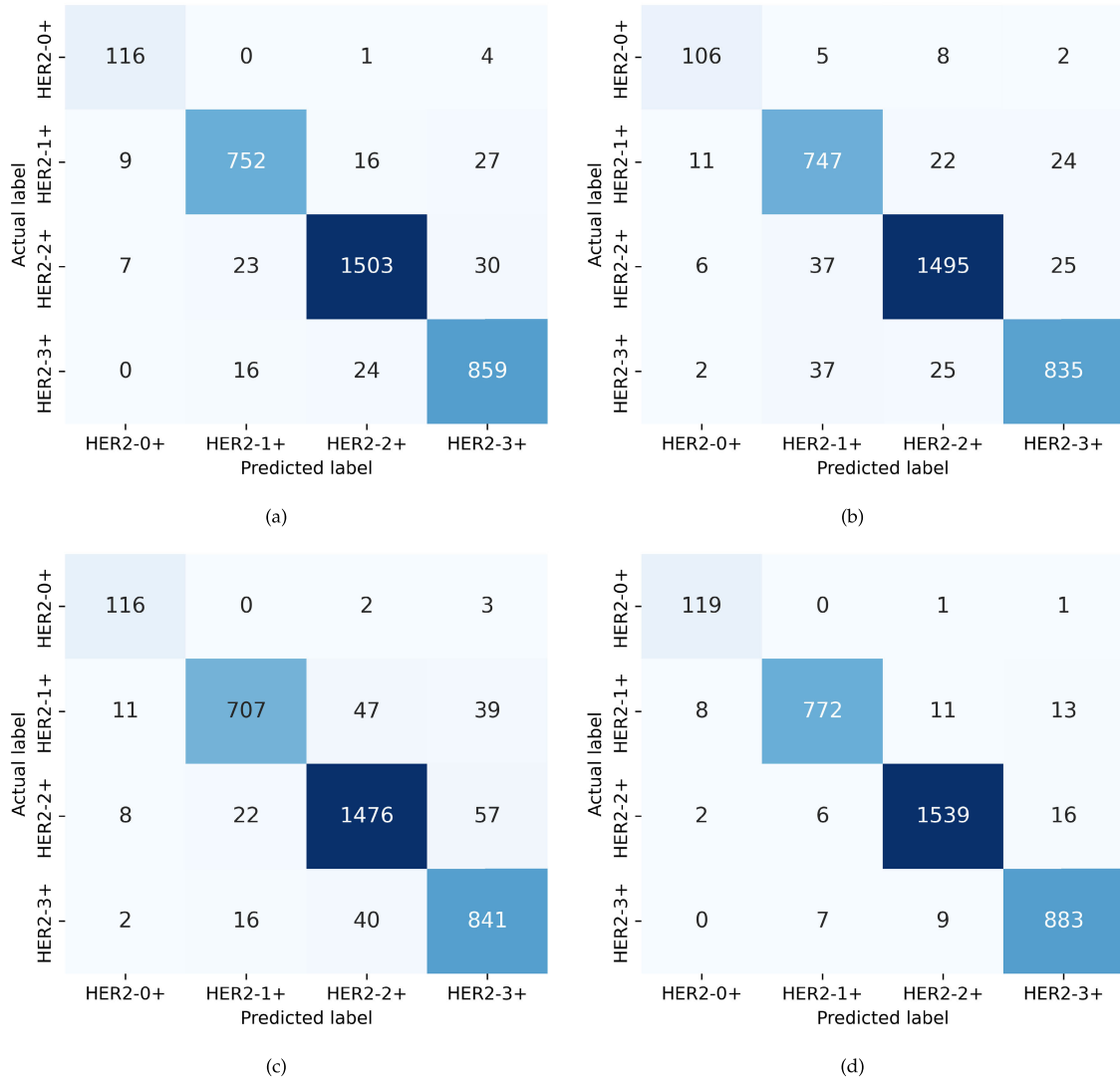


FIGURE 8. Confusion matrix of base classifiers and proposed method: (a) GoogleNet (b) DenseNet-201 (c) WideResNet-50 (d) Proposed method.

formulated as follows

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Here, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are the entries of the confusion matrix.

B. EXPERIMENTAL SETUP

The proposed method was implemented using PyTorch, an open-source framework based on Python. The training and testing were conducted on a system with access to the

workstation operated at CentOS Linux release 8.4.2105 system. The workstation is equipped with Intel(R) Xeon(R) Gold 6240R CPU @ 2.40Ghz, 32GB RAM, with a clock speed of 1000.730Mh.

C. ANALYSES OF BASE CLASSIFIERS

This section analyses the performance of the chosen three base classifiers GoogleNet, DenseNet-201, and WideResNet-50. Initially, these classifiers are fine-tuned and trained for 100 epochs using the HE-stained images from the BCI dataset. The hyper-parameters employed for training the models are discussed in Subsection III-C. The performances in terms of training and validation accuracy during the training phase of chosen base classifiers are plotted in Fig. 7, which demonstrate the progress of accuracy across each epoch. GoogleNet achieves a training and validation accuracy of 99.0% and 93.8%, respectively. It can be clearly noticed in

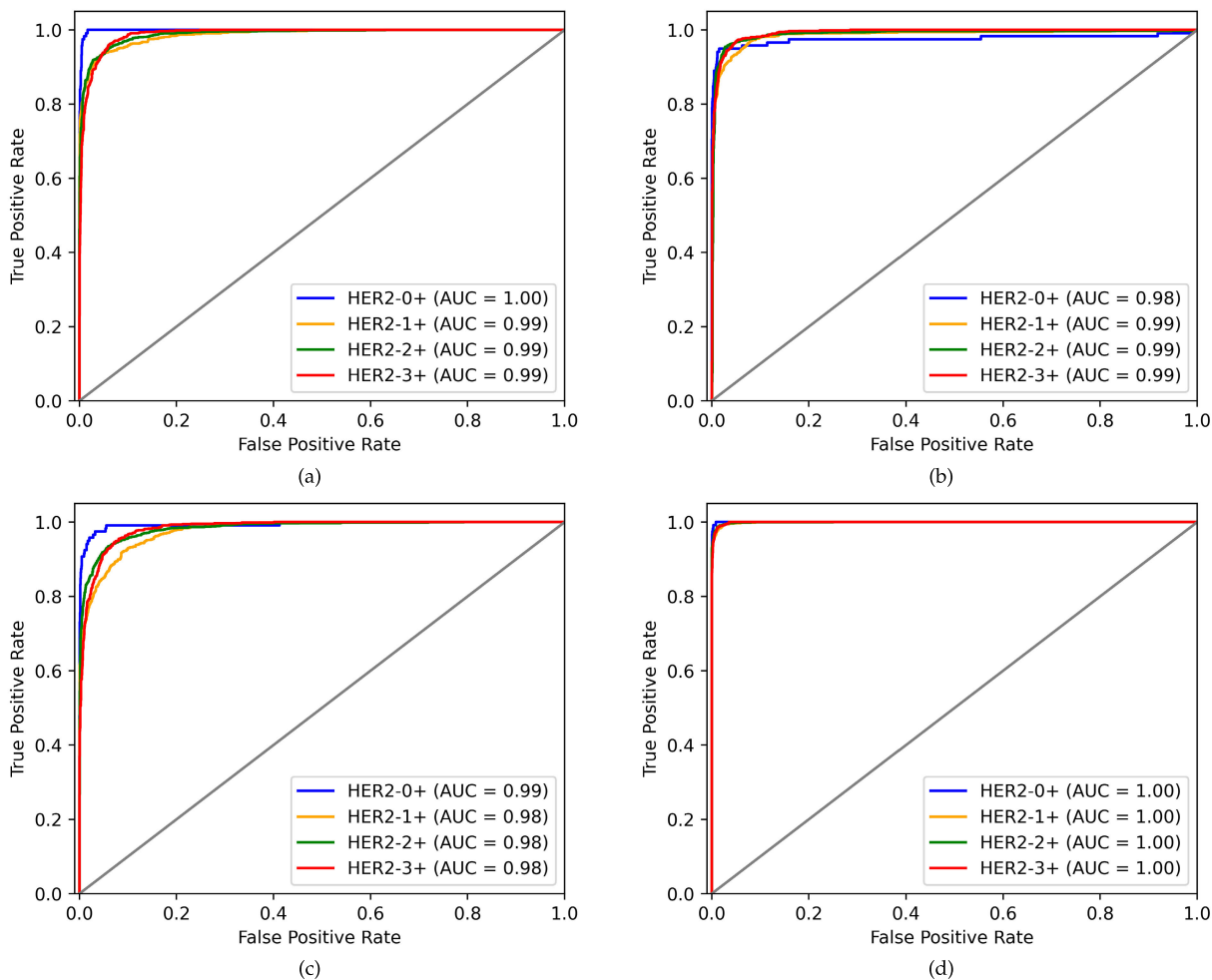


FIGURE 9. ROC curve and AUC values of: (a) GoogleNet (b) DenseNet-201 (c) WideResNet-50 (d) Proposed method.

Fig. 7 that the difference in the training and validation profile of GoogleNet indicates the issue of over-fitting whereas, in the case of the other two base classifiers, DenseNet-201 and WideResNet-50, although the issue of over-fitting is not observed, the accuracies achieved are not encouraging too. Both issues were addressed and resolved in the proposed method by introducing a novel weighted average ensemble algorithm.

D. ANALYSIS OF PROPOSED ENSEMBLE METHOD

We have proposed a novel and efficient ensemble algorithm to fuse the base classifiers at the decision level. Each base classifier can extract diverse and complementary visual features, enabling a more comprehensive representation of images from different perspectives. To enhance the generalisability as well as the prediction accuracy of the model, the three fine-tuned base classifiers are fused using a class-wise weighted average ensemble algorithm. As a result, an improved accuracy of 97.84% is obtained. The comparisons of all the performance measures of the proposed model with the base classifiers on the test set are given Table 3. Notably, the GoogleNet model exhibits better performance in accuracy,

precision, recall, and F1 score compared to the DenseNet-201 and WideResNet-50. However, the proposed class-wise ensemble algorithm utilizing an ensemble transfer learning strategy outperforms all individual base classifiers. This suggests the proposed approach exhibits promising potential for generalization compared to a standalone CNN classifier. The performance of the proposed model is evaluated using a confusion matrix, which provides more insight into model performance, errors, and weaknesses. The confusion matrices presented in Fig. 8 demonstrate the preeminent performance of the proposed algorithm on the test dataset. Subsequently, the comparison is carried out on the performance of the proposed class-wise ensemble method with the ensemble methodologies routinely explored in the literature, including average probability, weighted average probability, and majority voting. It is evident from Table 4 that, the proposed method outperforms all the traditionally employed ensemble methodologies. Finally, the receiver operating characteristic (ROC) comparison analysis and the AUC values are presented in Fig. 9. The proposed ensemble model has achieved an AUC value of 100% which illustrates the model’s superior performance.

TABLE 5. Comparison of proposed class-wise ensemble method with the other existing methods in the literature.

Model	Dataset	accuracy (%)	precision (%)	recall (%)	F1-Score (%)	AUC
HE-HER2Net [29]	BCI (HE)	87.01	87.73	87.00	87.11	91.00
DenseNet-201-Xception-SIE [30]	BCI (HE)	97.12	97.15	97.68	–	–
DenseNet-201-Xception-SIE [30]	BCI (IHC)	97.56	97.57	98.00	–	–
HAHNet [34]	BCI (HE)	93.65	93.67	92.46	93.66	99.00
ConvoHER2 [31]	BCI (HE)	85.10	–	–	–	–
ConvoHER2 [31]	BCI (IHC)	87.79	–	–	–	–
Her2Net [35]	Warwick (IHC)	98.33	96.64	96.79	96.71	–
Proposed Method	BCI (HE)	97.84	96.62	97.87	97.22	100

E. COMPARISON WITH THE OTHER EXISTING METHODS

The proposed class-wise ensemble method is compared with the existing deep transfer learning models in the literature. The outcome of this comparison analysis is presented in the Table 5. Among all the listed models with the BCI HE-stained image dataset, DenseNet-201-Xception-SIE [30] achieves the best result with an accuracy of 97.12%. Nevertheless, the performance lags by 0.82% compared to the proposed class-wise ensemble approach.

V. CONCLUSION AND FUTURE WORK

Breast cancer is a very lethal and dangerous disease among women. Early diagnosis of HER2 breast cancer with the help of deep learning methods can help patients and also medical experts make decisions and start an effective treatment. The conventional method for determining HER2 status in the medical domain typically involves the IHC test, followed by a detailed examination of IHC images by medical experts. However, this approach is both time-consuming and financially demanding for certain demographics. In contrast, the proposed automated framework leverages the HE images, bypassing the need for the IHC test to identify the four statuses (HER2-0+, HER2-1+, HER2-2+, HER2-3+) of HER2 over-expression effectively. This alternative methodology not only enhances classification accuracy but also expedites the entire diagnostic process significantly. Consequently, integrating the proposed algorithm into relevant hardware is simple and may lead to more efficient and streamlined early identification of HER2 over-expression.

Our future ventures will concentrate on the design and analysis of scanner-independent classification models. Also, one of our future works would focus on the spectral decomposition in terms of different color channels of histopathology images that may yield promising results for the effective classification of breast cancer histopathology microscopy images. Ultimately, conducting rigorous studies is imperative for advancing the early diagnosis of breast cancer and its subtypes. This can empower patients to mitigate their risks and make appropriate decisions about their health.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest.

FUNDING

This research did not receive any specific funds from any agencies in the public, commercial, or not-for-profit sectors.

ACKNOWLEDGMENT

This research was conducted without any external funding from funding agencies in the public, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTIONS

Pateel G. P.: Theoretical analysis, data curation, writing—original draft, and writing—review and editing. **Kedarnath Senapati:** Supervision, comparative analysis, writing—original draft, and writing—review and editing. **Abhishek Kumar Pandey:** Validation, writing—original draft, and writing—review and editing.

DATA AVAILABILITY

Dataset related to this article are taken from an open-source online data repository hosted at BCI.

REFERENCES

- [1] E. Arzanova and H. N. Mayrovitz, *The Epidemiology of Breast Cancer*. Brisbane, AU, USA: Exon Publications, 2022, pp. 1–19.
- [2] S. Arslan, X. Li, J. Schmidt, J. Hense, A. Gerales, C. Bass, K. Brown, A. Marcia, T. Dewhirst, P. Pandya, S. Singhal, D. Mehrotra, and P. Raharja-Liu, “Evaluation of a predictive method for the H&E-based molecular profiling of breast cancer with deep learning,” *bioRxiv*, Jan. 2022.
- [3] M. I. Nounou, F. ElAmrawy, N. Ahmed, K. Abdelraouf, S. Goda, and H. Syed-Sha-Qhattal, “Breast cancer: Conventional diagnosis and treatment modalities and recent patents and technologies,” *Breast Cancer, Basic Clin. Res.*, vol. 9, pp. 17–34, Sep. 2015.
- [4] N. Iqbal and N. Iqbal, “Human epidermal growth factor receptor 2 (HER2) in cancers: Overexpression and therapeutic implications,” *Mol. Biol. Int.*, vol. 2014, pp. 1–9, Sep. 2014.
- [5] I. Vaz-Luis, E. P. Winer, and N. U. Lin, “Human epidermal growth factor receptor-2-positive breast cancer: Does estrogen receptor status define two distinct subtypes?” *Ann. Oncol.*, vol. 24, no. 2, pp. 283–291, Feb. 2013.
- [6] Y. Che, F. Ren, X. Zhang, L. Cui, H. Wu, and Z. Zhao, “Immunohistochemical HER2 recognition and analysis of breast cancer based on deep learning,” *Diagnostics*, vol. 13, no. 2, p. 263, Jan. 2023.
- [7] ASCO. (2022). *Cancer.Net Editorial Board*. [Online]. Available: <https://www.cancer.net/cancer-types/breast-cancer/introduction/>

- [8] P. Pradhan, K. Köhler, S. Guo, O. Rosin, J. Popp, A. Niendorf, and T. Bocklitz, "Data fusion of histological and immunohistochemical image data for breast cancer diagnostics using transfer learning," in *Proc. 10th Int. Conf. Pattern Recognit. Appl. Methods*, 2021, pp. 495–506.
- [9] A. C. Wolff, M. E. H. Hammond, K. H. Allison, B. E. Harvey, P. B. Mangu, J. M. Bartlett, M. Bilous, I. O. Ellis, P. Fitzgibbons, and W. Hanna, "Human epidermal growth factor receptor 2 testing in breast cancer: American society of clinical oncology/college of American pathologists clinical practice guideline focused update," *Arch. Pathol. Lab. Med.*, vol. 142, no. 11, pp. 1364–1382, 2018.
- [10] R. Nahta and F. J. Esteva, "HER-2-targeted therapy: Lessons learned and future directions," *Clin. Cancer Res.*, vol. 9, no. 14, pp. 5078–5084, 2003.
- [11] F. Michas. (2018). *Number of Oncologists Per One Million People*. [Online]. Available: <https://www.statista.com/statistics/884711/oncologists-density-by-country-worldwide/>
- [12] M. S. A. Bychkov. (2023). *Global Pathologist Workforce*. [Online]. Available: <https://thepathologist.com/outside-the-lab/constant-demand-patchy-supply>
- [13] H. Jia, Y. Xia, Y. Song, D. Zhang, H. Huang, Y. Zhang, and W. Cai, "3D APA-Net: 3D adversarial pyramid anisotropic convolutional network for prostate segmentation in MR images," *IEEE Trans. Med. Imag.*, vol. 39, no. 2, pp. 447–457, Feb. 2020.
- [14] Y. Xie, Y. Xia, J. Zhang, Y. Song, D. Feng, M. Fulham, and W. Cai, "Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest CT," *IEEE Trans. Med. Imag.*, vol. 38, no. 4, pp. 991–1004, Apr. 2019.
- [15] K. Nazeri, A. Aminpour, and M. Ebrahimi, "Two-stage convolutional neural network for breast cancer histology image classification," in *Proc. 15th Int. Conf. Image Anal. Recognit. Povoá de Varzim, Portugal: Springer*, 2018, pp. 717–726.
- [16] S. Mehta, E. Mercan, J. Bartlett, D. Weaver, J. G. Elmore, and L. Shapiro, "Y-Net: Joint segmentation and classification for diagnosis of breast biopsy images," in *Proc. 21st Int. Conf. Med. Image Comput. Comput. Assist. Intervent. Granada, Spain: Springer*, 2018, pp. 893–901.
- [17] G. Litjens, C. I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. H. van de Kaa, P. Bult, B. van Ginneken, and J. van der Laak, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Sci. Rep.*, vol. 6, no. 1, pp. 1–11, May 2016.
- [18] A. Polónia, S. Campelos, A. Ribeiro, I. Aymore, D. Pinto, M. Biskup-Fruzyńska, R. S. Veiga, R. Canas-Marques, G. Aresta, T. Araújo, A. Campilho, S. Kwok, P. Aguiar, and C. Eloy, "Artificial intelligence improves the accuracy in histologic classification of breast lesions," *Amer. J. Clin. Pathol.*, vol. 155, no. 4, pp. 527–536, Mar. 2021.
- [19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
- [20] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, *arXiv:1605.07146*.
- [21] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
- [22] Y. Xie, J. Zhang, and Y. Xia, "Semi-supervised adversarial model for benign-malignant lung nodule classification on chest CT," *Med. Image Anal.*, vol. 57, pp. 237–248, Oct. 2019.
- [23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 84–90.
- [24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
- [25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
- [26] S. P. Oliveira, J. R. Pinto, T. Gonçalves, R. Canas-Marques, M.-J. Cardoso, H. P. Oliveira, and J. S. Cardoso, "Weakly-supervised classification of HER2 expression in breast cancer haematoxylin and eosin stained slides," *Appl. Sci.*, vol. 10, no. 14, p. 4728, Jul. 2020.
- [27] A. Maleki, M. Raahemi, and H. Nasiri, "Breast cancer diagnosis from histopathology images using deep neural network and XGBoost," *Biomed. Signal Process. Control*, vol. 86, Sep. 2023, Art. no. 105152.
- [28] G. Shamaï, Y. Binenbaum, R. Slossberg, I. Duek, Z. Gil, and R. Kimmel, "Artificial intelligence algorithms to assess hormonal status from tissue microarrays in patients with breast cancer," *JAMA Netw. Open*, vol. 2, no. 7, Jul. 2019, Art. no. e197700.
- [29] M. S. H. Shovon, M. J. Islam, M. N. A. K. Nabil, M. M. Molla, A. I. Jony, and M. F. Mridha, "Strategies for enhancing the multi-stage classification performances of HER2 breast cancer from hematoxylin and eosin images," *Diagnostics*, vol. 12, no. 11, p. 2825, Nov. 2022.
- [30] M. S. H. Shovon, M. F. Mridha, K. M. Hasib, S. Alfarhood, M. Safran, and D. Che, "Addressing uncertainty in imbalanced histopathology image classification of HER2 breast cancer: An interpretable ensemble approach with threshold filtered single instance evaluation (SIE)," *IEEE Access*, vol. 11, pp. 122238–122251, 2023.
- [31] M. F. Mridha, M. K. Morol, M. A. Ali, and M. S. H. Shovon, "ConvoHER2: A deep neural network for multi-stage classification of HER2 breast cancer," 2022, *arXiv:2211.10690*.
- [32] Y. Zheng, C. Li, X. Zhou, H. Chen, H. Xu, Y. Li, H. Zhang, X. Li, H. Sun, X. Huang, and M. Grzegorzczek, "Application of transfer learning and ensemble learning in image-level classification for breast histopathology," *Intell. Med.*, vol. 3, no. 2, pp. 115–128, May 2023.
- [33] S. Liu, C. Zhu, F. Xu, X. Jia, Z. Shi, and M. Jin, "BCI: Breast cancer immunohistochemical image generation through pyramid pix2pix," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1814–1823.
- [34] J. Wang, X. Zhu, K. Chen, L. Hao, and Y. Liu, "HAHNet: A convolutional neural network for HER2 status classification of breast cancer," *BMC Bioinf.*, vol. 24, no. 1, pp. 1471–2105, Sep. 2023.
- [35] M. Saha and C. Chakraborty, "Her2Net: A deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2189–2200, May 2018.



PATEEL G. P. received the B.E. degree in electronics and communication engineering and the M.Tech. degree in digital communication and networking engineering from Visvesvaraya Technological University, Belagavi, Karnataka, India, in 2018. He is currently pursuing the Ph.D. degree with the National Institute of Technology Karnataka, Surathkal, India. He was with the Department of Electronics and Communication Engineering, Visvesvaraya Technological University, until July 2022. His research interests include image processing, machine learning, and deep learning in medical applications. He is a Lifetime Associate Member of the Institution of Electronics and Telecommunication Engineers (IETE).



KEDARNATH SENAPATI received the Ph.D. degree in mathematics from the Institute of Mathematics and Applications, Bhubaneswar, in 2013. He was a Postdoctoral Researcher with IIT Bhubaneswar, in 2016. He is currently an Associate Professor with the Department of Mathematical and Computational Science, National Institute of Technology Karnataka, India. His research interests include the application of deep learning and machine learning in medical image analysis, signal processing, and spectral analysis.



ABHISHEK KUMAR PANDEY received the B.Tech. degree in electronics and instrumentation engineering from MAKAUT, Kolkata, in 2015, and the M.Tech. degree in engineering statistics from Cochin University of Science and Technology, Kerala, India, in 2020. He is currently pursuing the Ph.D. degree with the Department of Mathematical and Computational Sciences, National Institute of Technology Karnataka, Surathkal, India. His research interests include image processing, machine learning, and the application of deep learning in various medical imaging applications.

...